Working Memory-Driven Neural Networks with a Novel Knowledge Enhancement Paradigm for Implicit Discourse Relation Recognition

Fengyu Guo,[1,2,*] **Ruifang He,**[1,2,3,*†] **Jianwu Dang,**[1,2,4,†] **Jian Wang**[1,2]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China.
[2]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China.
[3]State Key Laboratory of Cognitive Intelligence, iFLYTEK, China.
[4]School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan.
{fengyuguo, rfhe, jian_wang}@tju.edu.cn, jdang@jaist.ac.jp.

Abstract

Recognizing implicit discourse relation is a challenging task in discourse analysis, which aims to understand and infer the latent relations between two discourse arguments, such as temporal, comparison. Most of the present models largely focus on learning-based methods that utilize only intra-sentence textual information to identify discourse relations, ignoring the wider contexts beyond the discourse. Moreover, people comprehend the meanings and the relations of discourses, heavily relying on their interconnected working memories (e.g., instant memory, long-term memory). Inspired by this, we propose a **K**nowledge-Enhanced **A**ttentive **N**eural **N**etwork (KANN) framework to address these issues. Specifically, it establishes a mutual attention matrix to capture the reciprocal information between two arguments, as instant memory. While implicitly stated knowledge in the arguments is retrieved from external knowledge source and encoded as inter-words semantic connection embeddings to further construct knowledge matrix, as long-term memory. We devise a novel paradigm with two ways by the collaboration of the memories to enrich the argument representation: 1) integrating the knowledge matrix into the mutual attention matrix, which implicitly maps knowledge into the process of capturing asymmetric interactions between two discourse arguments; 2) directly concatenating the argument representations and the semantic connection embeddings, which explicitly supplements knowledge to help discourse understanding. The experimental results on the PDTB also show that our KANN model is effective.

Introduction

Discourse relation describes how two adjacent text units (e.g. clauses, sentences, and larger sentence groups), called arguments, named *Arg*1 and *Arg*2, are connected semantically to one another. Implicit discourse relation recognition without explicit connectives (Pitler, Louis, and Nenkova 2009), which normally needs to be inferred from the specific context, is still a bottleneck of discourse analysis. It is also beneficial to many downstream NLP applications, such as machine translation, text summarization, information extraction and conversation system.

Previous studies mainly include 1) conventional discrete feature-based and 2) neural network-based models. The former adopt the artificially designed linguistic features and the complicated rules (Pitler, Louis, and Nenkova 2009; Rutherford and Xue 2014). However, implicit discourse relations are rooted in the semantics, which are difficult to be recognized from the surface features. Although neural network-based models obtain better argument representations and more precisely predict discourse relations (Zhang et al. 2015; Liu et al. 2016), they encode two discourse arguments without the interactive clues. The further approaches adopt different complicated neural networks (Lei et al. 2017; Guo et al. 2018) with attention mechanism (Liu and Li 2016; Cai and Zhao 2017), gate mechanism (Chen et al. 2016) or memory mechanism (Zhang, Xiong, and Su 2016) to mine the interactive information of argument pairs. However, they only focus on the intra-sentence textual information, neglecting the wider contexts beyond the discourse or relevant implicit clues.

In addition, the researchers in cognitive psychology argue that the ability of humans to remember and understand something depends not only on the different types of working memory (e.g., instant memory, long term memory), but also on the interconnections between them (Baddeley 2003), just as shown in Figure 1.

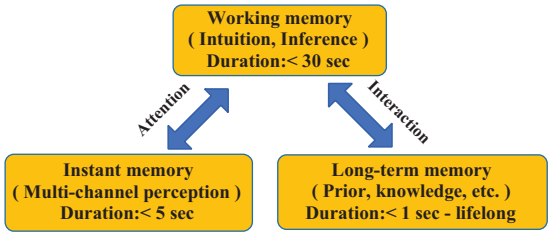


Figure 1: Working memory.

Intuitively, when annotating the meaning and the relations of discourse, people usually tend to capture the focused in-

*Equal contribution.
†Corresponding author.

formation as instant memory; meanwhile, the brain automatically wakes up the long-term memory, e.g., the relevant external knowledge. These two memories collaboratively promote to further comprehend the semantics of discourse. In fact, the existing methods encode knowledge in vector space, and directly concatenate the knowledge vectors to enhance the lexical or contextual representations of words in many tasks(Chen et al. 2018; Mihaylov and Frank 2018; Silva, Freitas, and Handschuh 2019), such as natural language inference, dialog generation and story completion. However, this paradigm does not fully fuse the text and its corresponding knowledge. Nowadays, there are many open knowledge bases (KBs)[1], such as WordNet(Miller 1995) storing semantic knowledge, ConceptNet (Speer, Chin, and Havasi 2017) storing commonsense knowledge and DBpedia(Lehmann et al. 2015) storing generic knowledge. In order to sufficiently comprehending the semantics of discourse relation, we choose the WordNet as our external knowledge.

Therefore, we propose **K**nowledge-Enhanced **A**ttentive **N**eural **N**etwork (KANN), a novel framework that is to imitate such human-like working memory for implicit discourse relation recognition. Especially, we capture the reciprocal information by mutual attention mechanism based on basic argument representation, for modeling the instant memory. Retrieved from an external source (i.e., WordNet), the relevant knowledge is harnessed to enrich the semantic understanding of discourse arguments, which explores a knowledge enhancement paradigm with implicit and explicit aspects. Finally, we integrate the argument and knowledge representations to improve the performance of this task.

In general, our main contributions are as follows:

- Propose a knowledge-enhanced attentive neural network for promoting comprehension of discourse arguments from the perspective of cognitive psychology;
- Imitate human-like working memory strategy: 1) exploit mutual attention mechanism to capture the interactive and significant information, as instant memory; 2) retrieve related knowledge from external source, as long-term memory;
- Devise a novel paradigm of knowledge enhancement with two ways: 1) mapping the knowledge matrix into attention matrix, an implicit way; 2) directly combining the inter-words semantic connection embeddings to obtain final representations, an explicit way;
- Experimental results on the PDTB show that our KANN model is effective, and the external knowledge is more significant when the size of data is restricted.

The Proposed Model

Implicit discourse relation recognition can be understood as a classification problem. However, not explicitly stated knowledge in discourse brings the difficulty of inferring discourse relation. The standard classifier with standard NLP techniques is not sufficient. In this section, we will explain how we integrate the external knowledge to make a final prediction. The proposed framework is shown in Figure 2.

[1]In this paper,we argue that Knowledge Base and Knowledge Graph can be interchanged in broad sense.

Instant Memory via Mutual Attention

Embedding Layer For the original representations of discourse arguments, we first associate each word w in the vocabulary with a vector representation $\boldsymbol{x}_w \in \mathbb{R}^d$ through an embedding lookup function, where d is the dimension of the embeddings. Since each argument is viewed as a sequence of word vectors, the arguments are expressed as:

$$Arg1 : [\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, ..., \boldsymbol{x}_{n_1}^1], \quad Arg2 : [\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, ..., \boldsymbol{x}_{n_2}^2].$$

where *Arg*1 has n_1 words and *Arg*2 has n_2 words.

Basic Argument Representation To represent the word in its context, we utilize a bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber 1997) to obtain a context-dependent hidden state at each position t of the sequence as:

$$\boldsymbol{h}_t = BiLSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}). \tag{1}$$

where $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}]$, and $\overrightarrow{\boldsymbol{h}_t}$, $\overleftarrow{\boldsymbol{h}_t}$ are the hidden states of the forward and backward layers which preserve the historical and future information. Therefore, *Arg*1 and *Arg*2 are encoded as $\boldsymbol{h}_i^1 = [\overrightarrow{\boldsymbol{h}_i^1}, \overleftarrow{\boldsymbol{h}_i^1}]$ and $\boldsymbol{h}_j^2 = [\overrightarrow{\boldsymbol{h}_j^2}, \overleftarrow{\boldsymbol{h}_j^2}]$, which are the intermediate states of i-th word in *Arg*1 and j-th word in *Arg*2 respectively, where $\boldsymbol{h}_i^1, \boldsymbol{h}_j^2 \in \mathbb{R}^{2d}$.

Separately encoding arguments by the basic BiLSTMs could not deeply reflect the asymmetric interactions between two arguments in a discourse relation. Here, the asymmetric interactions refer that different reading order of two arguments may lead to different focused information and relation decisions. Thus the interactions between two arguments are asymmetrical (Guo et al. 2018).

Mutual Attention Mutual attention (Santos et al. 2016) is aware of the input pair, in a way that semantic information from one argument can directly influence the other argument representation, and vice versa. The main idea is to automatically learn a similarity measure over the intermediate states in the argument pairs and use the similarity scores to compute attention vectors.

After obtaining the intermediate states of two arguments produced by BiLSTMs, we can obtain the matrices $\boldsymbol{R}^1 \in \mathbb{R}^{d \times n_1}$ and $\boldsymbol{R}^2 \in \mathbb{R}^{d \times n_2}$. And then we compute the matrix $\boldsymbol{G} \in \mathbb{R}^{n_1 \times n_2}$ as follows:

$$\boldsymbol{G} = \tanh(\boldsymbol{R}^{1^T} \boldsymbol{G}_0 \boldsymbol{R}^2). \tag{2}$$

where $\boldsymbol{G}_0 \in \mathbb{R}^{d \times d}$ is a matrix to be learned by the neural network and we employ *tanh* as activation function. And the element $\boldsymbol{G}_{i,j} \in \boldsymbol{G}$ is the pair-wise score of alignment between the hidden vectors of a word pair in two arguments. We call that the asymmetric interactions with lexical information reflecting the process of human-like instant memory to some extent.

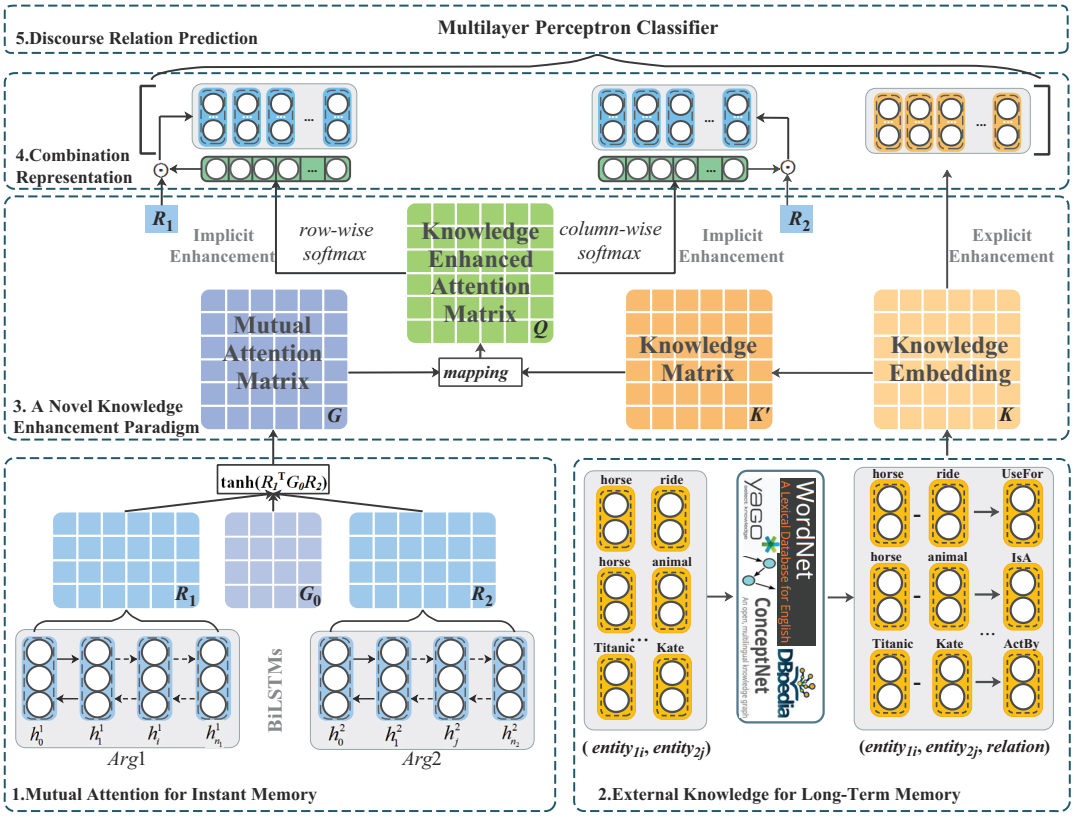


Figure 2: The overall architecture of our KANN model.

Long-Term Memory via External Knowledge

External Knowledge Retrieval Most open external knowledge sources are in the form of graphs. A knowledge graph is defined as a set of concepts connected by relations. In general, a fact of knowledge graph is represented as a triple $\boldsymbol{f}_i$=(*subject*, *relation*, *object*), such as "Barack Obama is the spouse of Michelle Obama" is represented as "Barack Obama, spouse, Michelle Obama", where "Barack Obama" and "Michelle Obama" are the subject and object, respectively, and "spouse" is their relation. Here, we utilize WordNet, a lexical database grouping nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), to capture the inter-words semantic connection embeddings[2] based on original discourse.

Given an instance (*Arg*1, *Arg*2), we traverse the two arguments to extract all word pairs in the arguments[3]. We obtain entity sets $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ from *Arg*1 and *Arg*2, respectively. And then $\boldsymbol{e}_{1i} \subset \boldsymbol{E}_1$, $\boldsymbol{e}_{2j} \subset \boldsymbol{E}_2$ are combined as entity pairs $(\boldsymbol{e}_{1i}, \boldsymbol{e}_{2j})$, which can be retrieved by looking up from WordNet.

Knowledge Embedding & Knowledge Matrix TransE (Bordes et al. 2013) achieves good results in knowledge representation, which models entity relations by interpreting them as translations in the low-dimensional space. We adopt TransE to capture semantic connection embedding between entity pairs in $(\boldsymbol{E}_1, \boldsymbol{E}_2)$. Specially, we exploit it to obtain the corresponding knowledge, which are trained on WordNet by the objective function $\boldsymbol{r} \approx \boldsymbol{h} - \boldsymbol{t}$, where $\boldsymbol{r}$ denotes the semantic connection embedding, $\boldsymbol{h}$ and $\boldsymbol{t}$ are the head and tail entity embeddings, respectively. If the i-th entity pair has multiple relations, the final semantic connection embedding is obtained by calculating an average for a weighted sum of these embeddings (pre-trained vectors) as Eq.(3), also called **knowledge embedding**.

$$\boldsymbol{r}_i = \frac{1}{m} \sum_{k=1}^{m} \boldsymbol{\mu}_k \cdot \boldsymbol{r}_k. \tag{3}$$

where the weight $\boldsymbol{\mu}_k$ is computed by:

$$\boldsymbol{\mu}_k = \frac{exp(\boldsymbol{r}_k)}{\sum_{j=1}^{m} exp(\boldsymbol{r}_j)}. \tag{4}$$

where m denotes the number of semantic connections in an entity pair.

Given the semantic connections between the entity pairs derived from the external knowledge, a **knowledge matrix**

[2]To distinguish discourse "relation", entity relation embedding is called semantic connection embedding.

[3]Here, entity pair below refers to word pair.

$\boldsymbol{K} \in \mathbb{R}^{n_1 \times n_2}$ is established, filled with the indicative function $\mathbb{I}(\boldsymbol{e}_{ij})$ as its elements.

$$\mathbb{I}(\boldsymbol{e}_{ij}) = \begin{cases} \boldsymbol{r}_i, & \text{if}(\boldsymbol{e}_{1i}, \boldsymbol{e}_{2j})\text{has a relation;} \\ 0, & \text{if}(\boldsymbol{e}_{1i}, \boldsymbol{e}_{2j})\text{has no relation.} \end{cases} \tag{5}$$

where the $\boldsymbol{e}_{ij}$ indicates the pair $(\boldsymbol{e}_{1i}, \boldsymbol{e}_{2j})$. And then we can obtain the relevant knowledge attention $\boldsymbol{K}' = f(\boldsymbol{K})$, where the function f is a non-linear function, such as *relu*, *tanh*.

This part reflects the process of human-like long-term memory to some extent. However, how they work together to promote the comprehension of discourse, we will explain it in detail.

A Novel Knowledge Enhancement Paradigm

Implicit Enhancement of Knowledge The entity pairs in two discourse arguments may benefit from external knowledge by mining their semantic connections. Given the mutual attention and the knowledge matrix, we map the knowledge matrix into attention matrix, which implicitly supplements the relevant knowledge into the process of capturing asymmetric interactions between arguments.

$$\boldsymbol{Q} = \boldsymbol{G} + \boldsymbol{K}'. \tag{6}$$

where $\boldsymbol{G}$ reflects the interactions between two arguments, and $\boldsymbol{K}'$ reflects the semantic connections between entity pairs in the arguments. Thus $\boldsymbol{Q} \in \mathbb{R}^{n_1 \times n_2}$ is an intra-sentence relation matrix integrated with the relevant knowledge.

We apply column-wise and row-wise pooling operations to generate the importance vectors. Since mean-pooling performs better than max-pooling in our experiments. Thus, we adopt mean-pooling operation by Eq.(7):

$$\begin{aligned} q_i^{\boldsymbol{R}^1} &= \text{mean}(\boldsymbol{Q}_{i,1}, \boldsymbol{Q}_{i,2}, ..., \boldsymbol{Q}_{i,n_2}), \\ q_j^{\boldsymbol{R}^2} &= \text{mean}(\boldsymbol{Q}_{1,j}, \boldsymbol{Q}_{2,j}, ..., \boldsymbol{Q}_{n_1,j}). \end{aligned} \tag{7}$$

where $q_i^{\boldsymbol{R}^1}$ represents an importance score for the context around the i-th word with external knowledge in *Arg*1 with regard to *Arg*2. Likewise, $q_j^{\boldsymbol{R}^2}$ represents an importance score for the context around the j-th word with external knowledge in *Arg*2 with regard to *Arg*1. So we can obtain the importance vectors of $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ as follows:

$$\begin{aligned} \boldsymbol{q}^{\boldsymbol{R}^1} &= [q_1^{\boldsymbol{R}_1}, q_2^{\boldsymbol{R}_1}, ..., q_{n_1}^{\boldsymbol{R}_1}]^T, \\ \boldsymbol{q}^{\boldsymbol{R}^2} &= [q_1^{\boldsymbol{R}_2}, q_2^{\boldsymbol{R}_2}, ..., q_{n_2}^{\boldsymbol{R}_2}]^T. \end{aligned} \tag{8}$$

Next, we utilize *softmax* function to transform these vectors $\boldsymbol{q}^{\boldsymbol{R}_1}$ and $\boldsymbol{q}^{\boldsymbol{R}_2}$ to obtain the knowledge-enhanced attention vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ shown in Eq.(9).

$$\boldsymbol{\alpha}_i = \frac{\exp(q_i^{\boldsymbol{R}_1})}{\sum_{j=1}^{n_1} \exp(q_j^{\boldsymbol{R}_1})}, \quad \boldsymbol{\beta}_i = \frac{\exp(q_i^{\boldsymbol{R}_2})}{\sum_{j=1}^{n_2} \exp(q_j^{\boldsymbol{R}_2})}. \tag{9}$$

Finally, the argument representations ($\boldsymbol{R}_{Arg1}$ and $\boldsymbol{R}_{Arg2}$) integrating the relevant deeper knowledge clues are calculated by Eq.(10), which not only captures the interactions between the arguments, but also reflects the construction of the long-term memory to some extent.

$$\boldsymbol{R}_{Arg1} = \boldsymbol{R}^1\boldsymbol{\alpha}, \qquad \boldsymbol{R}_{Arg2} = \boldsymbol{R}^2\boldsymbol{\beta}. \tag{10}$$

Explicit Enhancement of Knowledge The inter-words semantic connection embedding calculated by Eq.(3) is explicitly representing knowledge. It is directly concatenated to the argument representations, which is shown as Eq.(11).

Knowledge-Enriched Combination Representation

Although the representations computed by Eq.(10) incorporate the relevant knowledge implicitly, they do not adequately reflect the knowledge clues due to the lack of knowledge in intra-sentence as a whole (if only concatenating two arguments "[*Arg*1, *Arg*2]"). We exploit a composition layer to capture the context of whole discourse with semantic connection embeddings as follows:

$$\boldsymbol{R}_{total} = [\boldsymbol{R}_{Arg1}, \boldsymbol{R}_{Arg2}, \sum_{i=1}^{l} \boldsymbol{\nu}_i \boldsymbol{r}_i]. \tag{11}$$

where $\boldsymbol{\nu}_i$ is a soft-alignment weight which is similar to Eq.(9), l is the number of existing entity pairs in the arguments, and $\boldsymbol{r}_i$ is the semantic connection embedding in Eq.(3). Our model feeds the final representations into a classifier to determine the discourse relation. Here, we use the multilayer perceptron (MLP) classifier, which has one hidden layer with *tanh* activation and *softmax* output layer.

Discourse Relation Prediction

Given a training corpus which contains n instances $\{(\boldsymbol{x}, \boldsymbol{y})\}_{r=1}^{n}$, $(\boldsymbol{x}, \boldsymbol{y})$ denotes an argument pair and its relation label. We employ the cross-entropy loss to assess how well the predicted relation represents the real relation, defined as:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{j=1}^{C} \boldsymbol{y}_j \log(Pr(\hat{\boldsymbol{y}}_j)). \tag{12}$$

where $Pr(\hat{\boldsymbol{y}}_j)$ is the predicted probability of the j-th label, C is the number of relation class. To minimize the objective, we use stochastic gradient descent with the diagonal variant of AdaGrad with mini-batches.

Experiments

Data Preparation

Corpus We use the Penn Discourse TreeBank (PDTB) (Prasad et al. 2008), which is the largest hand-annotated discourse relation corpus annotated on 2,312 Wall Street Journal (WSJ) articles. Experiments are conducted on the four top-level classes as previous work (Rutherford and Xue 2014; Chen et al. 2016), namely, Comparison (Comp.), Contingency (Cont.), Expansion (Exp.) and Temporal (Temp.). Following the conventional data splitting, we use Section 2-20 as training set, Section 21-22 as test set, and Section 0-1 as development set. The relevant statistics of the four PDTB discourse relations are shown in Table 1.

Experimental Settings The 50-dimensional pre-trained word embeddings are provided by GloVe (Pennington, Socher, and Manning 2014), which are fixed during our model training. If there are words that are not in Glove, they

Relation	Train	Dev.	Test
Comparison	1842	393	144
Contingency	3139	610	266
Expansion	6658	1231	537
Temporal	579	83	55

Table 1: The statistics of implicit discourse relations in the PDTB.

are randomly generated in $[-1, 1]$. All the discourse arguments are padded to the same length of 50. Here, we do not present the details of tuning the hyper-parameters and only give their final settings as shown in Table 2.

Description	Value
The length of hidden states	50
Knowledge embedding size	300
Initial learning rate	0.001
Minibatch size	32
Dropout rate	0.5

Table 2: Hyper-parameters for our KANN model.

To evaluate our model, we adopt two kinds of experiment settings:1) the four-way classification to observe the overall performance; 2) the binary classification to solve the problem of unbalanced data in the training data, where one class is against the other three. We use an equal number of positive and negative instances in the training set in each class. The test set and development set retain the natural state.

Comparison Methods

We choose several competitive models as our baselines, including three aspects: argument representation, interaction and relevant knowledge.

1) Discourse Argument Representation

- **Liu2016**: (Liu and Li 2016) designed Neural Networks with Multi-level Attention to select the important words.
- **Rönnqvist2017**: (Rönnqvist and Chiarcos 2017) proposed an attention-based BiLSTM to model the arguments as a joint sequence.

2) Argument Pair Interaction

- **Chen2016**: (Chen et al. 2016) utilized a Gated Relevance Network (GRN) and incorporated both the linear and nonlinear interactions between word pairs.
- **Lei2017**: (Lei et al. 2017) adopted word-weighted averaging to encode argument representation, which could be incorporated with word pair information efficiently.

3) Relevant Knowledge

- **Lan2017**: (Lan et al. 2017) presented i) an attention-based neural network, which conducted the representation with interactions; and ii) a multi-task learning, which leveraged knowledge from auxiliary task to enhance the performance.

Model	Comp.	Cont.	Exp.	Temp.
Liu2016(2-level)	36.70	54.48	70.43	**38.84**
Liu2016(3-level)	39.86	53.69	69.71	37.61
Chen2016	40.17	54.76	-	31.32
Chen2016†	38.05	53.43	67.01	30.86
Lei2017	40.47	55.36	69.50	35.34
Lan2017	40.73	**58.96**	72.47	38.50
Lei2018	43.24	57.82	72.88	29.10
Our KANN	**43.92**	57.67	**73.45**	36.33

Table 3: Comparisons with the state-of-the-art models (%) on binary classification. "†" indicates that the experiment of model is replicated and the others are cited.

- **Lei2018**: (Lei et al. 2018) found semantic characteristics of each relation type and predicted the results by the specific properties[4].

In addition, we also use the three ablation models to compare with our KANN model.

- **LSTM**: we encode two discourse arguments by LSTMs, then concatenate them and feed to a MLP to predict the discourse relations.
- **BiLSTM**: on the basis of **LSTM**, we consider the bidirectional contextual information, and utilize BiLSTMs to encode two discourse arguments.
- **BiLSTM+Mutual Attention**: further, we construct the pair-wise matrix as mutual attention dynamically, and then integrate them to obtain the new argument representations (named **BMAN**).

The Overall Performance

Table 3 shows the F_1 scores of comparison models on binary classification, the observations are as follows:

1) On the whole, the performance of models based on argument representation is lower than that of the others. This could be caused by the parallel encoding of discourse arguments and neglecting the links between two arguments. The F_1 scores of knowledge-based models are higher than those of the others. It indicates that wider context is beneficial to this task and different knowledge may influence the prediction of different discourse relations.

2) For each relation, the F_1 scores of Temporal are the lowest in all models. This is reasonable since it accounts for the smallest number of instances (only 5%) in the corpus. With the increase of instance number in different discourse relations, the F_1 scores also rise. It proves that the corpus is also crucial to the task. And Lei2018 gains the worse F_1 score on Temporal, due to the lack of its specific linguistic properties in their manual analysis.

3) Our KANN model achieves the state-of-the-art F_1 scores on Comparison and Expansion relations than those of other models, which indicates the effectiveness of our knowledge enhancement paradigm. The reasons include two aspects: (i) some argument pairs may have confusing word pairs, which can be effectively mined by mutual attention;

[4]We argue that those properties as their linguistic knowledge.

Model	F1	Acc.
Liu2016(2-level)	46.29	57.17
Liu2016(3-level)	44.95	57.57
Rönnqvist2017†	39.51	49.22
Chen2016†	44.61	51.86
Lei2017	46.46	-
Lan2017	47.80	**57.39**
Lei2018	47.15	-
Our KANN	**47.90**	57.25

Table 4: Comparisons with the state-of-the-art models (%) on four-way classification. "†" indicates the same meaning as above.

(ii) only relying on the linguistic properties of discourse itself is insufficient. Some complex entity pairs which have indicative clues need to be further understood by leveraging external knowledge, especially, the synonyms and antonyms in WordNet are beneficial to the two discourse relations.

The results of four-way classification are shown in Table 4, and we make the following observations:

1) Argument representation-based models have the comparable F_1 scores with the methods based on pair interaction. This illustrates that properly representing arguments is as important as modeling the interactions. In addition, the F_1 score of Liu2016 (2-level) model is higher than that of three levels' (1.34%). It indicates that paying more attention may lead to the over-fitting problem.

2) The F_1 score of Lan2017 model is higher than that of other approaches, which achieves 1.41%, 1.34% improvements than Liu2016 (2-level) and Lei2017 respectively. It proves the clues that integrating the relevant knowledge into the representation is more important than only focusing the important information of the arguments themselves or their interactions.

3) The performance of our KANN model is comparable with that of Lan2017, both higher than Lei2018. The main reasons are: (i) multi-task applied in Lan2017 could obtain different kinds of auxiliary information from different corpora, and we also introduce external knowledge to enrich the semantic understanding of discourse; (ii) Lei2018 obtains the specific linguistic properties, which only focuses on the discourse self.

The Effectiveness of External Knowledge

We utilize three ablation methods to compare with our KANN model and obtain the following observations of the results in Table 5:

1) The performance of LSTM is the worst on all relations. Although BiLSTM captures more information than that of LSTM, the results are not very good. The reason is that separately encoding discourse argument by LSTM or BiLSTM ignores the local focused cues.

2) Compared with LSTM and BiLSTM, BMAN model achieves much better performance. It indicates that BMAN not only obtains the focused parts of discourse argument, but also captures the specific interaction clues by constructing the relevance of word pairs.

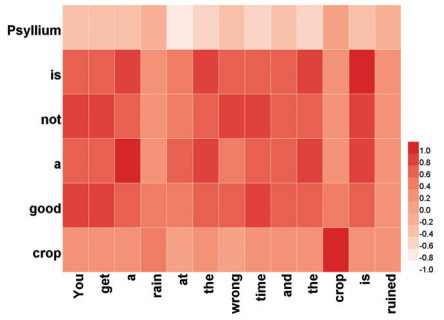
Model	F1	Acc.	
		Pre-trained	Randomly
LSTM	36.41	54.49	50.43
BiLSTM	36.54	55.31	52.31
BMAN	42.21	55.92	54.80
Our KANN	47.90	57.25	55.35

Table 5: The performance of ablation models with different setting on four-way classification.

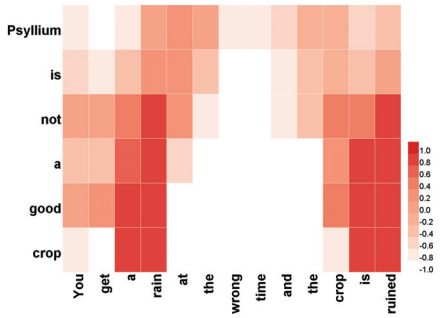
3) The accuracy (with pre-trained vectors) of our KANN model is slightly higher than the BMAN by 1.33%. We perform significance test for the improvement, and they are both significant under one-tailed t-test ($p < 0.05$). This manifests that some discourse arguments need rich knowledge to help the semantic understanding, and proves the effectiveness of the knowledge-enhanced attention module.

4) For the accuracy, all the ablation models with pre-trained embedding achieve much better performance than that of randomly initialized model. This reveals that the words associated with semantics could help the task. KANN (with randomly initialized vectors) has a comparable score with BMAN using the pre-trained vectors. It illustrates that the implicit and explicit utilization of external knowledge could effectively improve the comprehension of discourse.

Case Study



(a) Mutual attention.



(b) Knowledge-enhanced mutual attention.

Figure 3: A visualization of the attention matrices.

To demonstrate the validity of our external knowledge utilization, we visualize the heat maps of different attentions shown in Figure 3, which shows the attention matrices in an example. Every word accompanies with the various background colors. Darker patches denote higher correlations of word pairs. The example is listed below:

*Arg*1: *Psyllium's not a good crop.*

*Arg*2: *You get a rain at the wrong time and the crop is ruined.*

With respect to Figure 3a, we can observe the word pairs associating with "not", "good" are important context to determine the semantics which obtain much higher scores. It demonstrates the mutual attention could capture important parts of the arguments. However, the distribution of word pairs with higher scores is relatively average, which indicates that it is not enough to mine own information through such attention.

Figure 3b is as a comparison, the scores of word pairs are more prominent, which illustrates that integrating external knowledge makes the focused parts of arguments more clear. Meanwhile, it also reflects the characteristics of antonyms in WordNet, for example, the score of (good, wrong) is much lower.

Related Work

Traditional discrete feature-based methods for discourse relation recognition heavily rely on artificial and shallow features, such as POS, polarity, word position, etc. Recent neural network-based methods acquire the better performance, there are main three aspects as follows:

Discourse Argument Representation

The prerequisite of recognizing discourse relation is to have a good argument representation. Most previous researches exploit various neural networks, such as CNN, RNN, and hybrid models (Zhang et al. 2015; Qin, Zhang, and Zhao 2016a; Rutherford, Demberg, and Xue 2016) to encode discourse arguments as low-dimensional, dense and continuous representations. (Ji and Eisenstein 2015) integrated the linguistic features, including syntactic parsing and coreferent entity mentions into compositional distributed representations.

Though argument representation contains the high-level semantics, it does not embody emphasis during reading comprehension. And various attention mechanisms are used to reflect the emphasis on discourse arguments (Liu and Li 2016; Li, Li, and Chang 2016; Zhang, Xiong, and Su 2016). (Li, Li, and Chang 2016) exploited the hierarchical attention to capture the focus of different granularities. (Liu and Li 2016) imitated the repeated reading strategy, and proposed neural networks with multi-level attention (NNMA) to recognize discourse relations. However, these researches have not considered the reciprocal effects of two arguments at the beginning.

Argument Pair Interactions

Most studies tend to discover more semantic interactions between two arguments by complex neural networks (Chen et al. 2016; Qin, Zhang, and Zhao 2016b; Lan et al. 2017; Lei et al. 2017). (Chen et al. 2016) developed the deep neural architecture with a novel gated relevance network to capture semantic interactions between arguments. (Cai and Zhao 2017) generated discourse argument representations via pair-specified feature extraction. (Lei et al. 2017) conducted word interaction score to capture both linear and quadratic relation for argument representation. However, the studies only focused on intra-sentence information and ignored the external knowledge context.

Integration of External Knowledge

Most public available external knowledge sources are defined as a set of concepts connected by relations. Succeeding in many NLP tasks (Yang and Mitchell 2017; Chen et al. 2018; Mihaylov and Frank 2018) shows that external knowledge is effective for improving the performance of neural network-based models. (Yang and Mitchell 2017) incorporated knowledge directly into the LSTM cell state to improve event and entity extraction. (Chen et al. 2018) enriched neural network-based natural language inference (NLI) models with external knowledge in co-attention, local inference collection and inference components.

Different from the previous work, to our knowledge, we employ a paradigm with two ways of utilizing external knowledge to enrich the argument representations for the first time, which tries to break the inherent mode of traditional researches.

Conclusion and Future Work

Implicit discourse relation recognition demands sufficient understanding about the arguments from discourse itself and its relevant external knowledge. To this end, we propose to imitate the human-like working memory and exploit more comprehensive features through modeling the process of instant memory and long-term memory. Therefore, we design a novel neural **K**nowledge-Enhanced **A**ttentive **N**eural **N**etwork (KANN) framework. The new knowledge enhancement paradigm makes an effective fusion of discourse and its external knowledge. Thus, KANN can update the argument representations with corresponding knowledge, which can provide effective clues to identify discourse relations. Our experimental results on PDTB show that the proposed KANN model is effective.

In future work, we would like to explore different ways of integrating external knowledge into our task, and deeply investigate how to utilize the external knowledge inference for improving implicit discourse relation recognition.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Our work is supported by the National Natural Science Foundation of China (61976154), the Tianjin Natural Science Foundation (18JCYBJC15500), the National Key R&D Program of China (2018YFB1305200), the Tianjin Municipal Science and Technology Project (18ZXZNGX00330), and the Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK (CIOS-20190001).

References

Baddeley, A. 2003. Working memory: looking back and looking forward. *Nature reviews neuroscience* 4(10):829.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, 2787–2795.

Cai, D., and Zhao, H. 2017. Pair-aware neural sentence modeling for implicit discourse relation classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 458–466.

Chen, J.; Zhang, Q.; Liu, P.; Qiu, X.; and Huang, X. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th ACL*, 1726–1735.

Chen, Q.; Zhu, X.; Ling, Z.-H.; and Inkpen, D. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th ACL*, volume 1, 2406–2417.

Guo, F.; He, R.; Jin, D.; Dang, J.; Wang, L.; and Li, X. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th COLING*, 547–558.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Ji, Y., and Eisenstein, J. 2015. One vector is not enough: Entity-augmented distributional semantics for discourse relations. *Transactions of the Association for Computational Linguistics* 3:329–344.

Lan, M.; Wang, J.; Wu, Y.; Niu, Z.-Y.; and Wang, H. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 EMNLP*, 1299–1308.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6(2):167–195.

Lei, W.; Wang, X.; Liu, M.; Ilievski, I.; He, X.; and Kan, M.-Y. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *Proceedings of the 26th IJCAI*, 4026–4032.

Lei, W.; Xiang, Y.; Wang, Y.; Zhong, Q.; Liu, M.; and Kan, M.-Y. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, Q.; Li, T.; and Chang, B. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 EMNLP*, 362–371.

Liu, Y., and Li, S. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the EMNLP*, 1224–1233.

Liu, Y.; Li, S.; Zhang, X.; and Sui, Z. 2016. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2750–2756.

Mihaylov, T., and Frank, A. 2018. Knowledgeable reader: enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th ACL*, 821–823.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 EMNLP*, 1532–1543.

Pitler, E.; Louis, A.; and Nenkova, A. a. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of ACL and AFNLP*, 683–691.

Prasad, R.; Diesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.; and Webber, B. 2008. The penn discourse treebank 2.0. In *LREC*.

Qin, L.; Zhang, Z.; and Zhao, H. 2016a. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*, 70–77.

Qin, L.; Zhang, Z.; and Zhao, H. 2016b. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 EMNLP*, 2263–2270.

Rönnqvist, Samueland Schenk, N., and Chiarcos, C. 2017. A recurrent neural model with attention for the recognition of chinese implicit discourse relations. In *Proceedings of the 55th CL*, 256–262.

Rutherford, A., and Xue, N. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th EACL*, 645–654.

Rutherford, A. T.; Demberg, V.; and Xue, N. 2016. Neural network models for implicit discourse relation classification in english and chinese without surface features. *arXiv preprint arXiv:1606.01990*.

Santos, C. d.; Tan, M.; Xiang, B.; and Zhou, B. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Silva, V.; Freitas, A.; and Handschuh, S. 2019. Exploring knowledge graphs in an interpretable composite approach for text entailment. In *Thirty-Third AAAI conference on artificial intelligence*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 4444–4451.

Yang, B., and Mitchell, T. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *Proceedings of the 55th ACL*, volume 1, 1436–1446.

Zhang, B.; Su, J.; Xiong, D.; Lu, Y.; Duan, H.; and Yao, J. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 EMNLP*, 2230–2235.

Zhang, B.; Xiong, D.; and Su, J. 2016. Neural discourse relation recognition with semantic memory. *arXiv preprint arXiv:1603.03873*.